Briefings in Bioinformatics, 21(4), 2020, 1293–1301

doi: 10.1093/bib/bbz073
Advance Access Publication Date: 2 August 2019
Review article

Location deviations of DNA functional elements affected SNP mapping in the published databases and references

Hewei Zheng*, Xueying Zhao*, Hong Wang*, Yu Ding, Xiaoyan Lu, Guosi Zhang, Jiaxin Yang, Lianzong Wang, Haotian Zhang, Yu Bai, Jing Li, Jingqi Wu, Yongshuai Jiang and Liangde Xu

Corresponding authors: Hong Wang, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, P. R. China. Tel.: +86 (0) 451 8666 9617; Fax: +86 (0) 451 8666 9617. E-mail: wanghong84@ems.hrbmu.edu.cn; Yongshui Jiang, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, P. R. China. Tel.: +86 (0) 451 8666 9617; Fax: +86 (0) 451 8666 9617. E-mail: jiangyongshuai@gmail.com; Liangde Xu, School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou 325027, and Training Center for Students Innovation and Entrepreneurship Education, Harbin Medical University, Harbin 150081, P. R. China. Tel.: +86 (0) 577 8801 7534; Fax: +86 (0) 577 8801 7534. E-mail: xuld@eye.ac.cn

*These authors contributed equally to this work.

Hewei Zheng is a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. His main research interests are RNA molecular genetics and structural bioinformatics.
Xueying Zhao is a research assistant at the Institute of Hematology and Blood Diseases Hospital, Chinese Academy of Medical Sciences. Her main research interests are RNA molecular genetics and structural bioinformatics.
Hong Wang is an assistant professor at Harbin Medical University. Her research interests include complex disease bioinformatics and molecular genetics.
Yu Ding is a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. His main research interests are RNA and protein structural biology.
Xiaoyan Lu is a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. Her main research interests are RNA molecular genetics and structural bioinformatics.
Guosi Zhang a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. His main research interests are RNA molecular genetics and structural bioinformatics.
Jiaxin Yang is a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. Her main research interests are microbiological genomics and computational biology.
Lianzong Wang is a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. His main research interests are lncRNA regulatory bioinformatics and system biology.
Haotian Zhang is an undergraduate student jointly trained by Harbin Medical University and Wenzhou Medical University. His main research interests include bioinformatics algorithm and structural bioinformatics.
Yu Bai is a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. Her main research interests are RNA molecular genetics and structural bioinformatics.
Jing Li is a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. Her main research interests are RNA expression and molecular genetics.
Jingqi Wu is a graduate student jointly trained by Harbin Medical University and Wenzhou Medical University. Her main research interests are microbiology and bioinformatics.
Yongshuai Jiang is an associate professor at Harbin Medical University. His main research interests include complex disease bioinformatics and statistical genetics.
Liangde Xu is an associate professor at Eye Hospital of Wenzhou Medical University and Harbin Medical University. His main research interests include complex disease bioinformatics and molecular genetics.

Submitted: 4 March 2019; **Received (in revised form):** 24 May 2019

Abstract

The recent extensive application of next-generation sequencing has led to the rapid accumulation of multiple types of data for functional DNA elements. With the advent of precision medicine, the fine-mapping of risk loci based on these elements has become of paramount importance. In this study, we obtained the human reference genome (GRCh38) and the main DNA sequence elements, including protein-coding genes, miRNAs, lncRNAs and single nucleotide polymorphism flanking sequences, from different repositories. We then realigned these elements to identify their exact locations on the genome. Overall, 5%–20% of all sequence element locations deviated among databases, on the scale of kilobase-pair to megabase-pair. These deviations even affected the selection of genome-wide association study risk-associated genes. Our results implied that the location information for functional DNA elements may deviate among public databases. Researchers should take care when using cross-database sources and should perform pilot sequence alignments before element location-based studies.

Key words: functional elements; sequence alignment; location deviation; precision medicine

Introduction

The rapid development of high-throughput sequencing technologies and the associated extensive applications has driven biomedical research into the post-genomic era [1, 2]. Many large-scale international cooperative projects have surveyed the human genome [3], and the results of these studies have deepened our understanding of human physiology and pathology [4]. For example, the International HapMap Project (HapMap) has facilitated genome-wide association studies (GWASs) and human genome diversity research; these projects have identified a wide range of genes causing complex diseases and traits [5, 6]. In addition, the 1000 Genomes Project, a human genome map based on large samples from multiple populations, has further increased the depth and breadth of genome research [7, 8], while The Cancer Genome Atlas (TCGA) and other large projects are dedicated to genomic research focused on cancer and other major human diseases [9, 10]. These projects have explored genomic variation, modification, and expression in the etiology of disease and have created conditions for increasingly precise diagnosis and drug development [11, 12].

As genomic science has developed, hundreds of databases have been created, each oriented to different types of DNA sequence elements and/or specific research objectives [13]. The total amount of sequence data generated by large-scale genome projects has rapidly increased [14]. At present, DNA sequencing, RNA sequencing and various databases (e.g. GenBank, miRBase, ENCODE and dbSNP) and comprehensive platforms for sequence searching and downloading (e.g. UCSC and Ensembl) [15–18] provide convenient resources and technical support for functional genomics studies [19, 20].

Many recent studies have included several types of genomic functional elements [21], and investigations of interacting elements have become increasingly common. For example, studies have investigated how microRNAs (miRNAs) target mRNAs [22], how transcription factors and DNA interact [23], or how DNA and long non-coding RNAs (lncRNAs) regulate each other [24]. In addition, studies of polymorphisms and their effects continue to be important. Indeed, single nucleotide polymorphisms (SNPs) and single nucleotide variants (SNVs) are considered useful biomarkers of disease risk, as these factors affect the structures of coding and non-coding genes [25]. Precision medicine has promoted additional cross-omic approaches [26, 27]. As most studies of multiple DNA elements depend on location information acquired from sequence databases, accurate functional element locations within a genome are of great significance, especially for biological and medical studies beyond bioinformatics.

In pilot studies, we found that mainstream databases (e.g. GenBank, miRBase, ENCODE and dbSNP) differed slightly with respect to functional element locations, although most of these databases share the same reference genome [28–30]. These discrepancies may perhaps be ascribed to differences in the sequence alignment algorithms, or to data submission quality. However, these inconsistencies are inconvenient and may mislead researchers whose studies rely on the precise locations of functional elements. Therefore, it is necessary to systematically compare DNA element locations from different database sources.

Material and Methods

Human genome reference sequence acquisition and processing

The human reference genome sequence GRCh38, submitted by the Genome Reference Consortium (GRC) on 24 December 2013 [31], was downloaded from GenBank [15]. The reference genome was about 3.4 Gb long and included 22 autosomes, a pair of sex chromosomes and a mitochondrial DNA sequence. The mitochondrial DNA sequence was not included in our study. We retained the sequence information for each chromosome and stored these data in 24 independent files. We removed the headers, blanks, spaces and newline characters from each chromosome file and used the edited files to construct reference genome libraries.

Functional DNA element sequence acquisition and processing

Gene sequence acquisition and processing

We obtained data for 58 137 genes from Ensembl, including not only protein-coding genes but also partial pseudogenes and non-coding genes [32]. This gene set contained 19 786 clearly defined protein-coding genes. The remaining genes were remapped and used for the comparison of gene element locations among databases. Each location data item contained an Ensembl gene ID, associated chromosome, original genome location, gene type and complete sequence (including 5′- and 3′-UTRs). FASTA

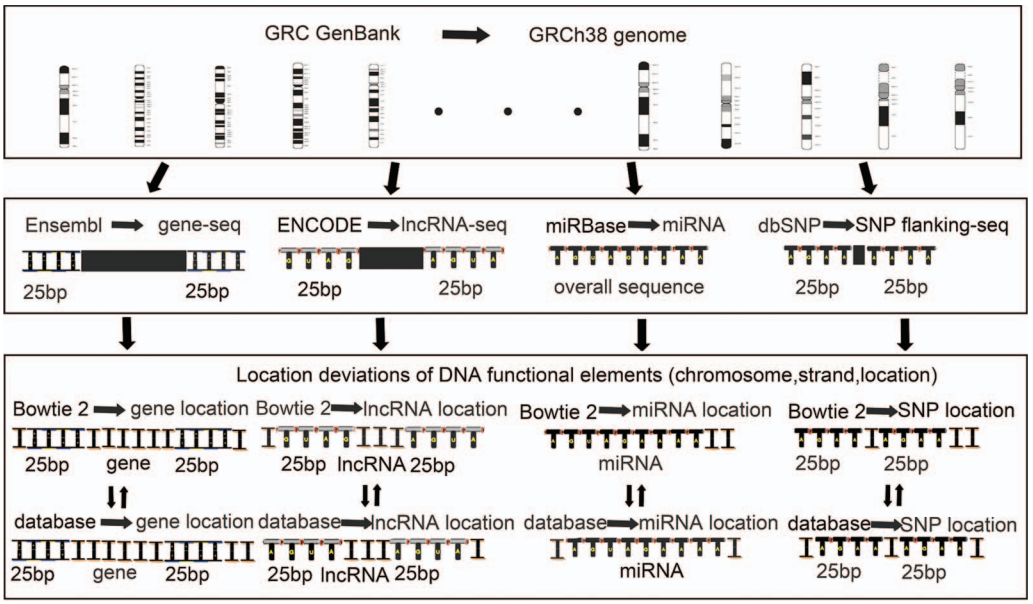

Figure 1. DNA fragment selection and alignment for different functional elements.

format sequence libraries were created by removing blank lines and newline characters.

Precursor and mature miRNA sequence acquisition and processing

We downloaded 1881 precursor miRNA sequences and 2588 mature miRNA sequences from miRBase (Release 22) [16]. We retained the miRBase ID, the original location and the sequence information but removed empty lines and breaks. We then transformed the RNA sequences into cDNA sequences and constructed FASTA format libraries.

LncRNA sequence acquisition and processing

LncRNA sequences were downloaded from ENCODE (Release 27) [17]: 27 908 transcripts (in which U was been replaced by T) encoding 15 778 lncRNA genes. LncRNA symbols, transcript IDs, original locations and sequences were retained and used to construct FASTA format libraries.

SNP-flanking sequence acquisition and processing

The SNP-flanking sequences of 324 709 505 SNPs were extracted from the NCBI dbSNP database (Build 150) [18]. SNP ID, original position, length of flanking sequence and allele information were retained. Each SNP allele was represented by the associated International Union of Pure and Applied Chemistry degenerate code. We removed whitespaces and blank lines, changed all lowercase letters into uppercase and then constructed FASTA format libraries using the modified data. We also downloaded SNP position information from HapMap and the 1000 Genomes Project to compare deviations across databases.

Functional DNA element alignment and mapping

Alignment fragment selection

To ensure alignment specificity, the alignment fragment length was set to 25 bp, which yielded a random probability of about 8.88×10^{-16}. For long sequence elements, such as protein-coding genes and lncRNAs, we used 25 bp from the head and tail of each sequence as the alignment fragments. For short sequences, such as precursor and mature miRNAs, the whole sequence was aligned. For SNPs, we aligned 25-bp fragments of the up- and down-stream sequences flanking each SNP loci. The details of alignment fragment selection are shown in Figure 1.

Functional element sequence alignment

After reference genome sequence library construction, we used Bowtie2 to align the functional elements [33, 34]. To generate precise alignments, the Bowtie2 parameters used were '-f –score -min L, 0, -0.3'. These settings ensured that the fragments were completely mapped on the reference genome. The alignment results indicated the associated chromosome, strand direction (+/−) and the start/end positions. Due to the broad distribution of DNA variation, we sometimes selected different alignment fragments and constructed a second alignment in order to increase location accuracy.

Functional element mapping

If both ends of a given sequence had consistent Bowtie2 alignments, the genome location of the elements could be calculated directly. The genome position of the element located on the '+' strand was defined as the 5′-25-bp start site to the 3′-25-bp end; the genome position of the element located on the '−' strand was the converse. The SNP position was defined as the larger of the two ends of the associated flanking sequence mapping results minus 1. The miRNA position was identical to the Bowtie2 result, and the mature miRNA was expected to be located within the host pre-miRNA sequence. If a functional element had location records in the associated source database, we compared these with the alignment results. For spliced lncRNA sequences from ENCODE, the coding DNA fragment location was recovered by the start and end sites of different transcripts.

Functional element location filtering and analysis

Based on the chromosome, strand, location and sequence length of each mapping result, we classified the obtained locations into

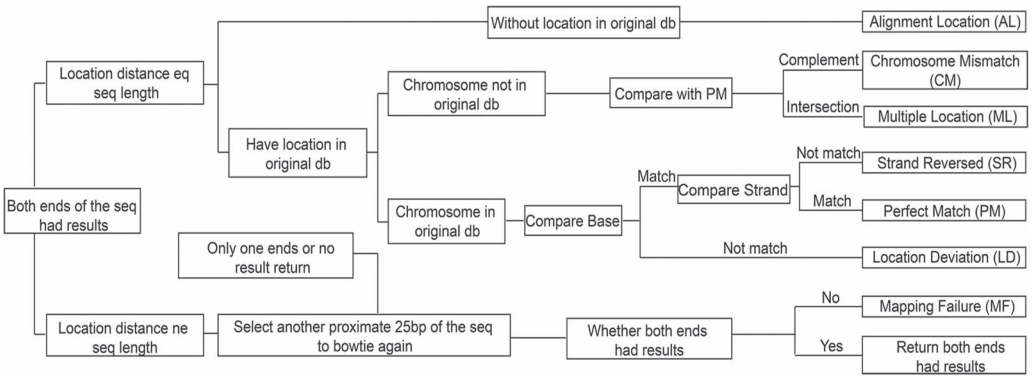

Figure 2. Functional element mapping and classification process.

seven groups: Perfect Match (PM), Location Deviation (LD), Strand Reversal (SR), Multiple Locations (ML), Chromosome Mismatch (CM), Alignment Location (AL) and Mapping Failure (MF). Results were classified as PMs when all the characters were consistent with the information in the original database. The LD elements were those where the chromosome and strand were consistent between our results and the original database, but the element location differed; location discrepancies inevitably affect subsequent location-based analyses. SR elements were mapped onto the opposite strand by our analysis versus the original database. ML elements were recovered by our analysis in more than one location on the genome. CM elements were mapped on to a different genome by our analysis, in comparison to the original dataset. This type of mismatch is important because *cis*-acting elements target the same chromosome or DNA molecule, while *trans*-acting elements target different chromosomes or DNA molecules. Thus, discrepancies in chromosomal locations among databases affect subsequent definitions of functional components. AL elements were those mapped to the genome by our analyses, but which were unmapped in the original database; these new genetic locations will facilitate future analyses. MF elements could not be mapped on the reference genome by our analysis.

However, single nucleotide peptides in genomic sequences might cause alignment failures at one or both fragment ends. Thus, for all elements classified as MFs, we extracted, re-aligned and re-classified a 25-bp fragment adjacent to the original fragment. All elements that were still not accurately mapped were considered the final MF group. All elements were classified into one, and only one, group. A detailed flowchart of the classification process is shown in Figure 2.

Multiple database location information comparison

Analysis of location deviation effects for GWAS

Genes associated with disease risk are primarily identified based on linkages with disease-associated SNPs. It is thus vital that the relative locations of SNPs and SNP-linked genes are accurate. We randomly selected GWAS-identified risk loci and their linked genes within 33 disease phenotypes in dbGAP (e.g. breast neoplasm, coronary artery disease, diabetes, hypertension, Parkinson's disease and prostatic neoplasm) [35], setting the risk threshold to 1.0×10^{-7}. We calculated the distances between the GWAS-identified SNPs and linked genes associated with disease and compared these to the distances identified in our analyses. We determined the potential effects of any location deviations on the GWAS results.

SNP location deviations in three key SNP repositories

We explored the degree of location deviation for a given functional element type among different data resources using SNPs as an example. First, we selected SNPs from dbSNP that were confirmed to be correctly located by our previous sequence alignment, as well as all the CEU (Europe) SNPs in HapMap and the 1000 Genomes Project. Using the SNP-ID, we compared the locations of the SNPs shared between HapMap and dbSNP, between HapMap and the 1000 Genomes Project and between dbSNP and the 1000 Genomes Project. We then analyzed the deviations in SNP location among these three data resources based on the pairwise comparisons.

Next, we extracted the identified MF SNPs from HapMap. We also obtained the minor allele frequencies (MAF) of CEU SNPs. We used a hypergeometric cumulative distribution to calculate the overrepresentation of rare SNPs (MAF < 0.01) in each subset as follows [36–38]:

$$P = 1\sum_{i=0}^{m-1}\frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}},$$

where N is the total number of SNPs in the HapMap CEU population; M is total number of MF SNPs; n is the total number of rare SNPs (MAF < 0.01) in the HapMap CEU population; and m is the number of MF SNPs with a MAF less than 0.01. Probability P described the relationship between SNP allele frequency and mapping failure.

Results

Gene sequence mapping and analysis

Among the 58 137 gene sequences from Ensembl, Bowtie2 alignment identified 52 408 bilateral sequences and 5729 unmapped sequences. We compared the sequence length of each bilateral sequence to the reference sequence and identified 1884 sequence length mismatches. We attempted to realign these 1884 sequences, as well as the 5729 unmapped sequences, using adjacent 25-bp fragments. Across all elements, 85.34% locations were PMs, and 9.85% were MFs (Figure 3A). The remaining element locations (4.81%) were corrected by further subdivision and were assigned to the ML, LD, CM or SR class. Figure 3B shows the classification of protein-coding genes after Bowtie2 alignment. These results indicate that, although the locations of most genes

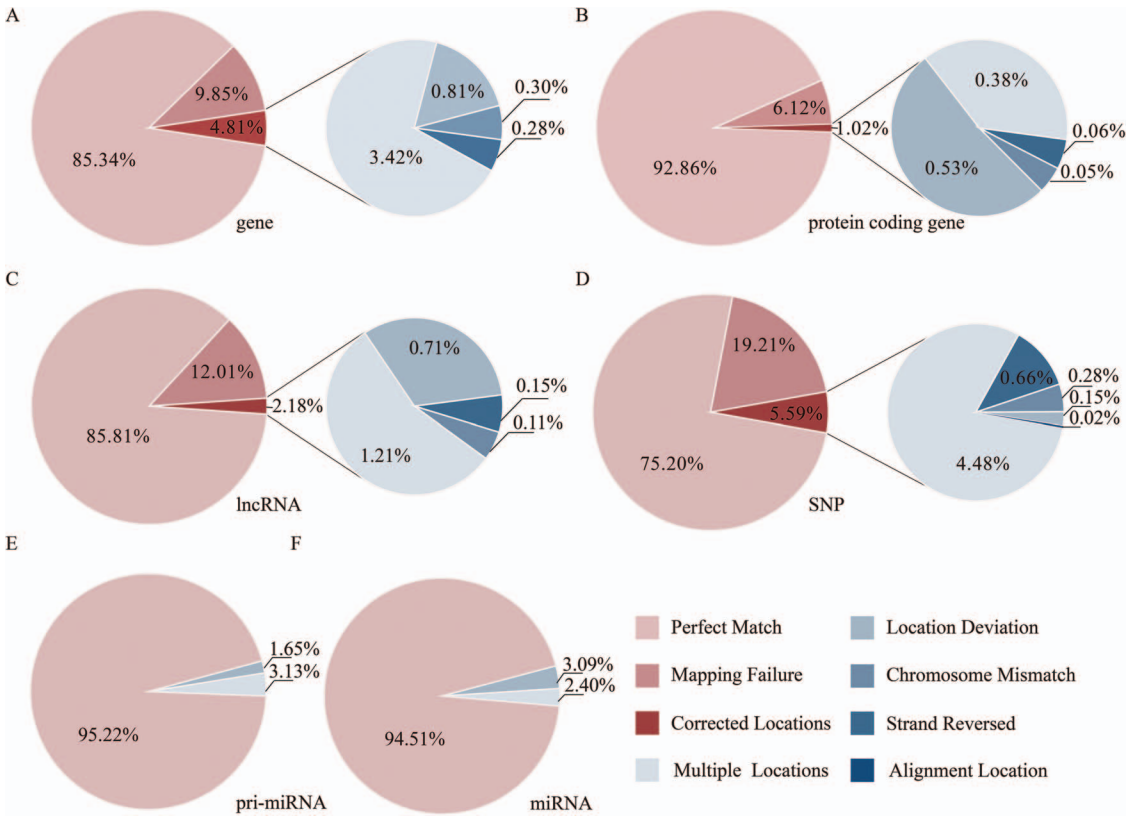

Figure 3. Functional element location situation distributions.

in most databases are accurate, some genes require remapping. These inconsistencies may have serious impacts on follow-up analyses.

Pre-miRNA and miRNA sequence mapping and analysis

We found that 95.22% of the 1881 pre-miRNA sequences obtained from miRBase were PMs, and ~4.78% were corrected to LDs or MLs (Figure 3E). None of the 1881 sequences was classified as SRs, ALs or MFs. Of the 2588 human miRNA sequences in miRBase, 94.51% were PMs, and ~5.49% were corrected to LDs or MLs (Figure 3F). None of the 2588 sequences was classified as SRs, ALs or MFs.

LncRNA transcript sequence mapping and analysis

Of the 27 908 lncRNA sequences in ENCODE, Bowtie2 alignment mapped 24 556 bilateral sequences, while 3352 sequences were MFs. We compared the Bowtie-derived lengths to the sequence lengths in ENCODE and identified 490 sequence length mismatches. We attempted to realign these 490 sequences, as well as the 3352 MFs, using additional 25-bp sequence fragments. Of the 3842 realigned sequences, 85.81% were PMs, 32 (0.11%) were CMs, 41 (0.15%) were SRs, 197 (0.71%) were LDs and 337 (1.21%) were MLs (Figure 3C).

SNP position mapping and analysis

The Bowtie2 alignment of the sequences flanking the 324 709 505 SNPs obtained from dbSNP recovered 324 613 893 bilateral alignments and 95 612 unmapped sequences. We compared the sequence length of each bilateral result derived from Bowtie2 to its reference sequence and identified 62 285 116 sequence length mismatches. We realigned these 62 285 116 sequences, as well as the 95 612 MFs. Of the realigned sequences, 75.20% were PMs, 19.21% were MFs and ~5.59% were corrected to LDs, MLs, CMs, ALs or SRs (Figure 3D).

Location deviations of risk SNPs and susceptible genes in GWAS

In the corrected gene and SNP location datasets, all risk-associated SNPs from dbGaP, in all disease phenotypes, were located on the same chromosome as their linked genes. In subsequent comparisons, the location deviation was considered '0' if the SNP was located exactly within the range of its linked gene. For all other SNPs, we calculated the shortest distance between the SNP and either end of the gene. These distances were divided into three categories: less than 100 kb, between 100 kb and 1 Mb and more than 1 Mb (Table 1). We found that 0%–31.67% of the risk-associated SNPs were located more than 100 kb from the original risk-associated gene, which exceeds the extent of linkage disequilibrium blocks.

Location deviation of SNPs among data resources

The results of the pairwise comparisons among HapMap, dbSNP and the 1000 Genomes Project were sorted by deviation degree. We then constructed a 100% stacked column chart of the number of SNP location deviations between any pair of databases (Figure 4). We found that many (36.36%) of the SNPs shared

Table 1. Location deviation of GWAS risk SNPs and linkage genes in dbGaP. The first column of the table shows disease phenotypes in dbGaP. Other columns represent the percentage of different location deviation distribution

Phenotype	0 (%)	x ≤ 100 Kb (%)	100 Kb–1 Mb (%)	x > 1 Mb (%)
Anatomy category	44.76	23.57	27.62	4.05
Bacterial infections and mycoses	8.91	89.11	1.98	0.00
Behavior and behavior mechanisms	51.04	21.68	23.78	3.50
Behavioral disciplines and activities	58.82	17.65	23.53	0.00
Biological marker	56.24	25.00	15.63	3.13
Body weight and measure	54.17	24.11	18.30	3.42
Cardiovascular disease	54.34	31.76	12.90	1.00
Chemical and drugs category	59.63	25.75	12.78	1.84
Congenital, hereditary, and neonatal diseases and abnormalities	46.30	31.48	22.22	0.00
Diagnostic techniques and procedures	57.74	23.10	16.54	2.62
Digestives system disease	42.34	40.05	15.82	1.79
Disorder of environment origin	52.63	26.32	21.05	0.00
Endocrine system disease	51.13	40.81	7.30	0.76
Eye diseases	57.44	37.44	5.12	0.00
Female urogenital diseases and pregnancy complications	46.66	36.00	16.67	0.67
Hemic and lymphatic disease	56.63	32.53	9.64	1.20
Immune system disease	42.19	50.90	6.33	0.58
Laboratory techniques and procedure	62.16	29.34	7.72	0.78
Male urogenital disease	41.56	38.68	18.93	0.83
Mental disorders	54.18	26.69	17.13	2.00
Musculoskeletal diseases	44.32	36.36	18.18	1.14
Neoplasms	55.45	30.51	12.83	1.21
Nervous system diseases	46.65	43.10	9.62	0.63
Nutritional and metabolic disease	54.10	24.46	19.88	1.57
Otorhinolaryngologic disease	45.00	45.00	10.00	0.00
Parasitic diseases	0.00	100.00	0.00	0.00
Pathological conditions, signs and symptoms	55.17	22.13	20.40	2.30
Physical examination	54.50	22.97	19.22	3.31
Physical phenomena and processes	58.21	12.69	27.61	1.49
Respiratory tract diseases	46.95	45.22	7.83	0.00
Skin and connective tissue diseases	41.06	52.21	6.37	0.36
Stomatognathic diseases	40.00	45.71	14.29	0.00
Virus diseases	80.00	20.00	0.00	0.00
Others	54.27	36.11	8.53	1.09

between HapMap and dbSNP had location deviations >1 Mb, as did many of the SNPs shared between HapMap and the 1000 Genomes Project (Figure 4C and A). These discrepancies may have been due to the different release of the reference genome used by HapMap. However, even though dbSNP and the 1000 Genomes Project use the same version of the reference genome, many SNP location deviations were still identified between these databases (Figure 4B).

The hypergeometric cumulative distribution test showed that, in general, the MF SNPs were overrepresented among the SNPs with low MAF (MAF < 0.01; $P < 3.45 \times 10^{-6}$). This indicated that lower frequency SNPs were more likely to be false-positive errors. This may be because lower frequency SNPs/SNVs in a given population are sometimes presented as singletons, resulting in insufficient research and an absence of double-hit experimental confirmation.

Discussion

The accurate location of genomic functional elements, especially across multiple element types, is important for biomedical research [39]. In this study, inconsistent location information was identified by the realignment of functional element sequences from various databases to the reference genome.

For gene sequences from Ensembl, the overall location accuracy was 85.34%. Notably, protein-coding gene location consistency was 92.86%, probably due to the depth and breadth of the available global research on coding genes. Nevertheless, the observed 7.14% of coding genes with inconsistent locations was symptomatic of the potential problems across mainstream databases.

The location accuracy of miRNAs and pre-miRNAs in miRBase was about 94.51% and 95.22%, respectively. This high accuracy might be due to the short length of these sequences, as well as their relative rarity. Compared to other RNA types, miRNAs have received much research attention, which also has improved location reliability. Sustained attention to fewer than 2000 miRNAs has led to a relatively full understanding, which has improved reliability. As the average lengths of miRNAs and pre-miRNAs are only 21 bp and 80 bp, respectively, these sequences can be easily mapped onto the genome. However, random alignment errors may also be produced.

Only 85.81% of the lncRNAs from ENCODE were classified PMs, far less than the percentages of PM protein-coding genes and pre-mRNAs. This difference might be explained by the rapid increase in the numbers of lncRNAs recognized in recent years due to next-generation sequencing (NGS). In addition, effective classification, nomenclature and sequence submission criteria are lacking for lncRNAs, obviously impacting data

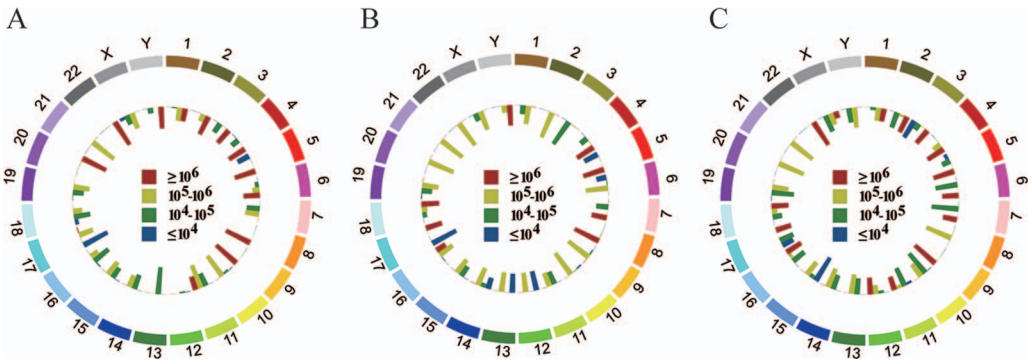

Figure 4. Location deviations of SNP among dbSNP, 1000 Genomes Project and HapMap. A. Location comparison between HapMap and 1000 Genomes Project; B. Location comparison between dbSNP and 1000 Genomes Project; C. Location comparison between dbSNP and HapMap. Four squares with different colors in each core represent the location deviation of SNP between diverse resources. The bar of inside circle shows the proportion of location deviation on each chromosome. The external circle of each subgraph shows 24 chromosomes.

standardization. Here, we obtained reliable locations for some lncRNAs using sequence alignments and mapping. These locations might be useful for lncRNA functional analyses and biological annotation.

The location accuracy for SNPs from dbSNP was only about 75.20%. In addition, relatively few SNPs were PMs. These location problems might have been due to the low frequency of some SNPs in the human population. Therefore, we performed a hypergeometric cumulative distribution test on the MF and low-frequency SNPs in HapMap. We found that MF SNPs are overrepresented among the low-frequency SNPs (MAF < 0.01 and $P < 3.45 \times 10^{-6}$), suggesting that SNP frequency affects alignment-derived locations.

dbSNP is a comprehensive database of variation, which accepts data from public projects and private research organizations. More recently, HapMap and the 1000 Genomes Project have been developed. Most groups submitting to these databases use NGS, which generates many SNPs and SNVs [40]. However, the diverse methods of data submission used by the various groups have given rise to many data consistency problems, such as a lack of data validation and low-quality data.

When comparing and analyzing data from different sources, we identified a certain degree of deviation both between GWAS disease-risk SNPs and their linked genes and between SNPs and their associated lncRNAs. We also found data source differences in the genomic locations of the same types of functional DNA elements. Some of these deviations were caused by differences in the release version of the reference genomes used by the databases when locating functional elements. However, even when the reference genome version was the same, relatively minor deviations were still observed, due to differences in other important steps, such as data preprocessing, alignment strategy and parameter setting.

In our study, we download reference genome sequences from UCSC. After fragment alignment using Bowtie2, the genome location of each element was calculated. We categorized 222 genes, 103 lncRNAs and 58 PCGs as LDs and categorized 207 genes, 101 lncRNAs and 55 PCGs as PMs. This difference might be due to the constant updating of the database. We also identified 86 genes, 16 lncRNAs and 14 PCGs at different locations on the same chromosome and strand, consistent with our previous results (e.g. ENSG00000273610: chr1, +, 21 987 481–21 987 777 and chr1, +, 22 010 650–22 010 946). We also identified extra matching positions during the mapping process (e.g. ENSG00000268993: chr1, −, 121 142 051–121 142 438; chr1, −, 206 160 889–206 161 276; and chr1, −, 143 929 995–143 930 382). The LD gene, PCG and lncRNA sequences were compared to the reference genome using BLAST. BLAST showed that 105 genes, 51 lncRNAs and 34 PCGs were PMs, while 86 genes, 9 lncRNAs and 9 PCGs had different locations on the same chromosome and strand. These latter genes, lncRNAs and PCGs were also identified by our Bowtie2 analysis. However, since Bowtie2 identified an additional seven lncRNAs and five PCGs, Bowtie2 might be more suitable for mapping short fragments to reference genomes.

In our study, we found that many participants in an SNP–gene linkage relationship in GWAS are disturbed by gene and SNP location deviations in the dbGaP database. To universalize our study, we used the GWAS catalog database to perform a further study, by comparing the mapping results from the dbGaP and GWAS catalog. In this process, we mainly focused on prostatic neoplasms, breast neoplasms, colorectal neoplasms, lung neoplasms, ovarian neoplasms and stomach neoplasms, which have been identified as SNPs with the most risk. In the dbGaP database, we found 460 SNPs and 383 genes, while in the GWAS catalog database, there were 581 SNPs and 594 genes. For the mapping results, we found that there were 18 and 19 genes that couldn't match the accuracy position for the dbGaP and GWAS catalog, respectively. The results showed that there were some errors in gene mapping due to gene location deviation in both the risk SNP genotype and phenotype databases. By comparing with the result of the UCSC and Ensembl gene location analysis, we infer that some of the information stored in the database has been revised with the update of data release, while information has not been updated in time for other databases.

As described above, for the six diseases (prostatic neoplasms, breast neoplasms, colorectal neoplasms, lung neoplasms, ovarian neoplasms and stomach neoplasms), we performed the SNP and gene alignment, from the dbGap and GWAS catalog, and found many mismatches between risk SNPs and their capture genes (Supplementary Table S1). For example, the original location of MIR4752 (ENSG00000264703) in Ensembl is on chromosome 19, +, 54 282 109–54 282 180, but our alignments show that its real location is chromosome 19, +, 54 629 392–54 629 463, and the horizontal shift is 347 283 bp. In the prostatic neoplasm's genome-wide association study, the identified risk SNP rs103294 (chromosome 19, 54 293 995) is adjacent to the original location of MIR4752 (with the distance of 11 815 bp). However, the realignment location of MIR4752 is obviously beyond (335 397 bp) the scope of linkage disequilibrium to rs103294. This means that the candidate for prostate cancer risk microRNA mir-4752 may

not be the real risk factor, based on the theory of gene association study. The similar samples also exist in breast neoplasms, colorectal neoplasms and ovarian neoplasms.

Our results suggest that the functional DNA element sequence and location information deviates among public databases, reminding researchers to be careful when using cross-database sources. In particular, sequences should be realigned first, especially in element location-based studies.

Key Points

- Relatively unneglectable discrepancies/deviations are noted when performing a comprehensive analysis of location alignments in functional DNA elements across several sequence repositories.
- A degree of up to 20% location deviations is detected in different scales of kilobase- to megabase-pair location alignment analysis.
- The GWAS candidate gene mapping is affected by SNP location deviations in a large part of disease and phenotype studies.
- A sequence alignment work strongly suggested before disease phenotype studies, especially for element location-based studies.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Authors' contributions

H.Z., X.Z. and H.W. carried out the bioinformatics analysis and drafted the manuscript. Y.D., X.L., G.Z. and J.Y. performed the bioinformatics analysis, drafted the manuscript and participated in the manuscript revision process. L.W., H.Z., Y.B., J.L. and J.W. plotted the main figures in the context. H.W., Y.J. and L.X. presented this idea, designed the experiment and guided the whole research process. All authors read and approved the final manuscript.

Acknowledgements

The author wish to thank all members and the support from the Training Center for Students Innovation and Entrepreneurship Education, Harbin Medical University, Harbin 150081, China.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 31801098, 91746113 and 31501062), the Research Projects of Education Department of Heilongjiang Province (grant numbers 1254G041 and 12511273), the Research Project of Health Department of Heilongjiang Province (grant number 2011-249), the Harbin Science and Technology Bureau project (grant number RC2013QN004112), the Fundamental Research Funds for the Provincial Universities (grant number 2017JCZX50) and the Internal Fund Project of Eye Hospital of Wenzhou Medical University (grant number YJGG20181001).

References

1. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet* 2008;**24**:133–41.
2. sequencing MV N-G. The genome jigsaw. *Nature* 2013;**501**:263–8.
3. Green ED, Watson JD, Collins FS. Human Genome Project: twenty-five years of big biology. *Nature* 2015;**526**:29–31.
4. Zhou M, Hu L, Zhang Z, *et al*. Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol Ther Nucleic Acids* 2018;**12**:518–29.
5. Couzin J. Human genome. HapMap launched with pledges of $100 million. *Science* 2002;**298**:941–2.
6. International HapMap C. The International HapMap Project. *Nature* 2003;**426**:789–96.
7. Siva N. 1000 Genomes Project. *Nat Biotechnol* 2008;**26**:256.
8. Kuehn BM. 1000 Genomes Project finds substantial genetic variation among populations. *JAMA* 2012;**308**:2322, 2325.
9. Kim HS, Minna JD, White MA. GWAS meets TCGA to illuminate mechanisms of cancer predisposition. *Cell* 2013;**152**:387–9.
10. Zhou M, Zhao H, Wang X, *et al*. Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Brief Bioinform* 2018.
11. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 2015;**19**:A68–77.
12. Ding Y, Wang H, Zheng H, *et al*. Evaluation of drug efficacy based on the spatial position comparison of drug-target interaction centers. *Brief Bioinform* 2019.
13. Pandey A, Lewitter F. Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem Sci* 1999;**24**:276–80.
14. O'Rawe JA, Ferson S, Lyon GJ. Accounting for uncertainty in DNA sequencing data. *Trends Genet* 2015;**31**:61–6.
15. Burks C, Cinkosky MJ, Fischer WM, *et al*. GenBank. *Nucleic Acids Res* 1992;**20**(Suppl):2065–9.
16. Griffiths-Jones S, Grocock RJ, van Dongen S, *et al*. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006;**34**:D140–4.
17. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;**306**:636–40.
18. Smigielski EM, Sirotkin K, Ward M, *et al*. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 2000;**28**:352–5.
19. Hubbard T, Barker D, Birney E, *et al*. The Ensembl genome database project. *Nucleic Acids Res* 2002;**30**:38–41.
20. Karolchik D, Baertsch R, Diekhans M, *et al*. The UCSC Genome Browser Database. *Nucleic Acids Res* 2003;**31**:51–4.
21. Bernstein BE, Kellis M. Large-scale discovery and validation of functional elements in the human genome. *Genome Biol* 2005;**6**:312.
22. Yang C, Sun C, Liang X, *et al*. Integrative analysis of microRNA and mRNA expression profiles in non-small-cell lung cancer. *Cancer Gene Ther* 2016;**23**:90–7.
23. Arlt A, Sebens S, Krebs S, *et al*. Inhibition of the Nrf2 transcription factor by the alkaloid trigonelline renders pancreatic cancer cells more susceptible to apoptosis through decreased proteasomal gene expression and proteasome activity. *Oncogene* 2013;**32**:4825–35.

24. Zhou M, Diao Z, Yue X, *et al*. Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer. *Oncotarget* 2016;**7**:56383–94.
25. Lam ET, Bracci PM, Holly EA, *et al*. Mitochondrial DNA sequence variation and risk of pancreatic cancer. *Cancer Res* 2012;**72**:686–95.
26. Kaiser JBIOMEDICINE. NIH opens precision medicine study to nation. *Science* 2015;**349**:1433.
27. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. *Nature* 2015;**526**:336–42.
28. Wang H, Lu X, Chen F, *et al*. Landscape of SNPs-mediated lncRNA structural variations and their implication in human complex diseases. *Brief Bioinform* 2018.
29. Wang H, Zheng H, Wang C, *et al*. Insight into HOTAIR structural features and functions as landing pads for transcription regulation proteins. *Biochem Biophys Res Commun* 2017;**485**:679–85.
30. Wang C, Wang L, Ding Y, *et al*. LncRNA structural characteristics in epigenetic regulation. *Int J Mol Sci* 2017;**18**.
31. Guo Y, Dai Y, Yu H, *et al*. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* 2017;**109**:83–90.
32. Birney E, Andrews TD, Bevan P, *et al*. An overview of Ensembl. *Genome Res* 2004;**14**:925–8.
33. Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics*, *Chapter 11:Unit* 2010;**11**:17.
34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.
35. Tryka KA, Hao L, Sturcke A, *et al*. NCBI's database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* 2014;**42**: D975–9.
36. Mazza T, Mazzoccoli G, Fusilli C, *et al*. Multifaceted enrichment analysis of RNA-RNA crosstalk reveals cooperating micro-societies in human colorectal cancer. *Nucleic Acids Res* 2016;**44**:4025–36.
37. Vossen DM, Verhagen CVM, Grenman R, *et al*. Role of variant allele fraction and rare SNP filtering to improve cellular DNA repair endpoint association. *PLoS One* 2018;**13**: e0206632.
38. Chen SF, Chao YL, Shen YC, *et al*. Resequencing and association study of the NFKB activating protein-like gene (NKAPL) in schizophrenia. *Schizophr Res* 2014;**157**:169–74.
39. Lai WK, Buck MJ. ArchAlign: coordinate-free chromatin alignment reveals novel architectures. *Genome Biol* 2010;**11**: R126.
40. de Vries PS, Sabater-Lleal M, Chasman DI, *et al*. Comparison of HapMap and 1000 genomes reference panels in a large-scale genome-wide association study. *PLoS One* 2017; **12**:e0167742.